Deep Learning based Automatic Extraction of Student Performance from Gazette Assessment Data

Rohan Nikam[1,*], *Ritesh* Pardeshi[2,**], *Yash* Patel[3,***], and *Ekta* Sarda[4,****]

[1]Department of Computer Engineering, Ramrao Adik Institute of Technology, India
[2]Department of Computer Engineering, Ramrao Adik Institute of Technology, India
[3]Department of Computer Engineering, Ramrao Adik Institute of Technology, India
[4]Department of Computer Engineering, Ramrao Adik Institute of Technology, India

Abstract. Everyday millions of files are generated worldwide containing humongous amounts of data in an unstructured format. Most of us come across at least one new document every week, which tells the large volume of data associated with documents. All the data in these documents is in unstructured format which makes it difficult for further processing. The extraction of data from this documents still remains largely a manual effort resulting in higher processing time. A system that could extract the required fields from documents and store them in a structured format automatically will be of much significance. In this paper, we have described an approach for extracting the data from Exam Result Gazette document and then storing it in a CSV file. Mask RCNN model having a backbone of ResNeXt-101-32x8d and Feature Pyramid Network(FPN) has been hypertuned for detecting the required fields. Then PyTesseract Optical Character Recognition System has been used for extracting the data from detected fields. Our proposed system is trained on custom data set created by us and then evaluated on test data to extract the required fields. The overall accuracy of our system is 98.69%. The results indicate that the system could be used for efficiently extracting the required fields from given exam result gazette document.

1 Introduction

Documents are the primary tools to accumulate information in each and every domain [1]. Few of the domains which heavily rely on documents are Healthcare domain having Medical records, Medical test reports of patients, Banking domain consisting mainly of Credit card statements, Account statements, Tech domain consisting of Resumes, Education domain consisting of Registration forms, and Marksheets. At present, manual work is mostly involved in processing these documents. For extracting the data from these documents and storing it in database organisations employ a separate team of individuals. So this becomes a time consuming and expensive process [2]. All the documents present in every domain have valuable data locked in them. If this data is extracted and stored in a structured format then it could be utilized for fast and efficient searching. Various analytic algorithms could be applied to data for business processing to get deeper insights out of it. A system that can automatically extract all these data can drastically improve the efficiency of many business processes by avoiding error-prone manual operations. In this paper, we used a college Exam Result Gazette Document or we can use any documents to extract the features, for example, as shown in Figure 1. We used the Mask RCNN model [3] with ResNeXt-101-32x8d as the backbone [4] and the Feature Pyramid Network (FPN)[5]. To detect the relevant fields from Exam Result Gazette Document, we performed hyper-parameter tuning of the model. COCO annotation format [6] is used as input to the model. In order to train the model from accessible Exam Result Gazette Document, we created unique training data in COCO annotation format.

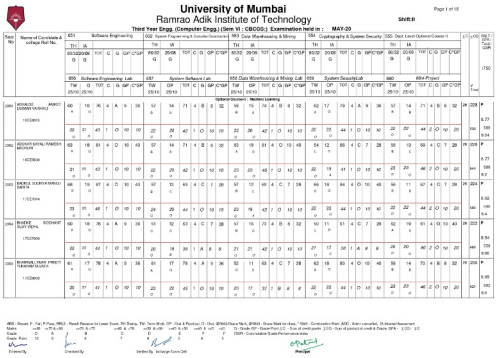
University of Mumbai
Ramrao Adik Institute of Technology

Figure 1. Exam Result Gazette Document

The detected fields given by the model are extracted and are then passed to Optical Character Recognition system (OCR) to extract the text from detected image and is then stored in a CSV file.

*e-mail: rohananilnikam1@gmail.com
**e-mail: riteshnleo@gmail.com
***e-mail: yashpatel7025@gmail.com
****e-mail: ekta.sarda@rait.ac.in

2 Literature Survey

B. Majumder [7], defined representation learning to identify target fields. In a document, not all the data present is of use. So hence a model was proposed which studies the relationship and relevancy between labels and data. The limitation of this model was that it was generalized to only specific fields like numbers in the document. **T. Denk [8]**, proposed a Combined Grid-based approach with BERT-based text encodings. BERTgrid, based on Chargrid, is a text-based grid that embeds vectors, making its structure and semantics accessible to the neural processing network. The inconvenience of that particular article was that it was more expensive computationally because the grid approach treated raw pixels.

V. Sunder [9], defined two techniques, The first is neural learning which uses a pre-trained deep learning model for reading and converts document images into structured form by adding a predefined database schema and the second is reusable logic programs for the extraction of entities from a document using the entities and primitive links identified through neural learning to synthesize. A template-free solution has been learned to detect pre-printed text and enter text/handwriting and to predict pairwise relationships. The pre-printed text and input text lines were detected via a convolutional network method. Then functionality from the detection network is integrated to classify potential connections in a semantic way. **B. Davis [10]**, proposed combination method exceeds heuristic regulations and that visual characteristic is essential to achieving high accuracy.

S. Paliwal [11], for both table detection and structure recognitions, describe deep learning end-to-end models. The model takes advantage of the interdependence between the two tasks of table detection and table structure recognition, which are divided between table and column regions. This is followed by semantic rule-based row extraction from the table subregions identified. Content-based, noise-resistant, and generalized machine-learning approach for unseen document formats. This addressed the task of removing important fields from a variety of document formats that are potentially unseen. SYPHT – a machine learning solution for the extraction of field documents – was introduced in this paper **X. Holt [12]**. SYPHT combines OCR, heuristic filtering, and the supervised document ranking model in order to predict the field level so that image quality, skew, orientation, and content layout changes can be robust.

X. Zhao [13], used the grid text in Convolution Neural networks where text is embedded into the document as a feature containing both semantic meaning and spatial distribution. After studying this paper it is learned that semantic feature extraction needs to be improved along with taking into consideration the image level features. **A. Katti [14]**, a new type of text representation has been introduced which preserves a document's 2D layout. It proposed a novel paradigm for processing and understanding structured documents. Instead of serializing a document into a 1D text, the proposed method, named chargrid, preserves the spatial structure of the document by representing it as a sparse 2D grid of characters. The model predicts a segmentation mask with pixel-level labels and objects bounding boxes to group multiple instances of the same class.The chargrid paradigm is applied on an information extraction task from invoices and demonstrates that this method is superior to both, state-of-the-art NLP algorithms as well as computer vision algorithms.

G. Sehgal [15], developed a framework that recognizes handwritten and printed text, removes noisy effects, identifies documents and visual entities, such as tables, linens, and boxes. One limitation of this is that if there is background noise in the image, it does not function well when recognizing handwritten text. **R.Palm [16]** describes a system that used RNNs and LSTM to capture the context of data in the document. On the basis of the context of words in the document, the model tries to generalize onto unseen templates. The model presented in this paper does not take into consideration the spatial layout of the document and follows a left to right order to extract key, values pairs.

3 Proposed Work

After doing the literature survey we have observed that there are some limitations of existing systems. Some existing systems that we have studied only work well with numerical data and some systems are not able to extract the data accurately which is in tabular format.

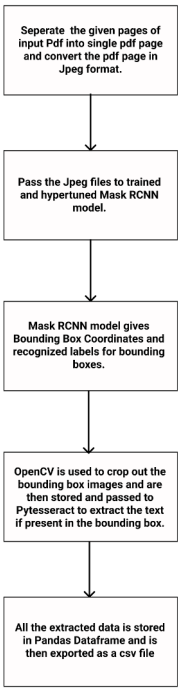

Figure 2. Flow Diagram

In the Figure 2, as we have mentioned that our system can work with textual as well as numerical data present in the document. As shown, the system first takes an input of Exam Result Gazette PDF, separates each page of the PDF,

converts it into a JPEG format and then passes it to the Mask RCNN model having a backbone of ResNeXt-101-32x8d and Feature Pyramid Network(FPN). Hyper-tuned Mask RCNN model gives the bounding box coordinates and recognized labels for each corresponding bounding box.

OpenCV is used to crop out the bounding box images and then these images are passed onto rightly configured Optical Character Recognition System PyTesseract to extract the text if present in the bounding box. Finally, all the extracted data is stored in Pandas Dataframe and exported as a CSV file.

3.1 Techniques

For creating the data for training open source Dataturks Annotation tool has been used. Dataturks annotation tool gives annotation in COCO Format which stores annotation in JSON Format. In the document, there are various fields like Name, Roll Number, Total Marks, Average Pointer, etc. Bounding boxes are drawn along these fields.

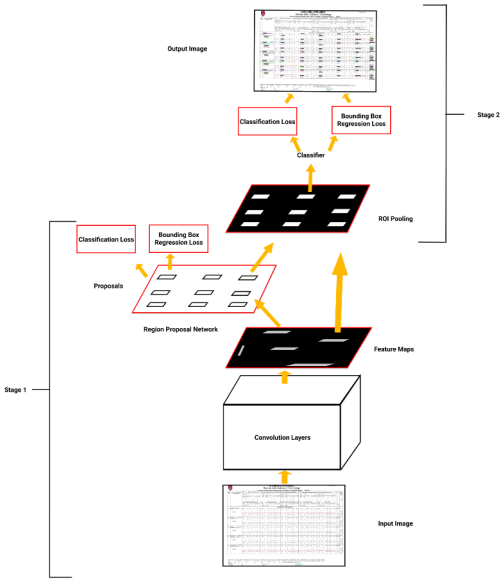

Figure 3. Working of Model

Mask RCNN model having a backbone of ResNeXt-101-32x8d and Feature Pyramid Network(FPN) has been used for detecting and recognizing various fields present in the document. Figure 3 shows the working of the model. Hyper-parameter Tuning has been done in Mask RCNN for increasing the accuracy of detecting and recognizing various fields present in the Exam Result Gazette document. Mask RCNN model works in two stages, the first stage is the region proposal stage, and the second stage is a network to predict most probable bounding boxes, classes of bounding boxes and mask for each bounding box. In the first stage, two networks are used backbone network and the region proposal network. Both these networks' backbone and region proposals run once per image to give a set of region proposals. The backbone network consists of a CNN to generate multiple Regions Of Interest using a lightweight binary classifier. For this, the model uses nine anchor boxes over the entire image. The classifier returns a score on the basis of whether an object is present in the proposed boxes or not. On the basis of threshold scores, some proposed boxes are selected and are warp into a fixed dimension. In the second stage, this fixed dimension of features is passed to a fully connected layer to make classification using the Softmax function. Warped features are also passed to a Mask Classifier. Mask Classifier is a CNN model which outputs a Mask for each Region Of Interest. In both, the stages of Mask RCNN Classification Loss and Bounding-Box Regression Loss have been used.

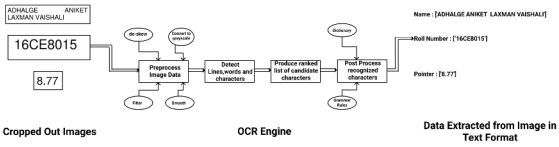

Figure 4. Working of OCR

After predicting the boxes from the document, the whole image, which is passed to the optical character recognition system (OCR), is cut into these boxes. The general working of OCR is shown in Figure 4. OCR scans each of the cropped images pixels by pixel to predict the text sent inside the cropped image. Then all the data recognized with OCR technique is exported to a CSV file.

3.2 Implementation Details

The Dataturks annotation tool was used to create the first custom data set of documents. Dataturks annotation tool after annotating the images gives the output in COCO format. For implementing the whole model, the Google Colaboratory platform has been used. Graphical Processing Unit(GPU) available on Google Colaboratory has been used For training the hyper-tuned model. PyTorch and Detectron2 Framework have been used.

Mask RCNN model having a backbone of ResNeXt-101-32x8d and Feature Pyramid Network(FPN) has been used for detecting and recognizing various fields present in the document. Model is trained using a custom training set created by Dataturks annotation tool. Various hyper-parameters in the model are adjusted in order to get high accuracy in detecting and predicting the fields of the documents. Once the fields are detected in the document, the coordinates of bounding boxes enclosing the detected fields and their respective labels are given as output by the model. These coordinates along with the image of gazette document are passed to OpenCV library for cropping out the bounding boxes from whole image. All the cropped images are stored according to their predicted labels in respective folders in order to pass to Optical Character Recognition System.

For extracting the text from the cropped images, the PyTesseract Optical Character Recognition Open Source library has been used. To correctly extract text from cropped images, various PyTesseract configurations were tested, and the one with the highest accuracy was chosen

as the final configuration. Different configurations have been used for extracting the numerical data like marks and textual data like name. Data of students with only unique roll numbers are stored. This data is stored in a Pandas DataFrame and is then exported as a CSV file.

4 Result and Analysis

The hyper-tuned Mask RCNN model was trained on Exam Result Gazette Document consisting of 25 pages, each page having 5 records. The trained model was evaluated on Exam Result Gazette Document consisting of 15 pages, each page having 5 records. Each record has 15 fields, such as Name, Roll Number, Pointer, Average Pointer, Total Marks, Subject1 Marks, Subject2 Marks, Subject3 Marks, Subject4 Marks, Subject5 Marks, Lab1 Marks, Lab2 Marks, Lab3 Marks, Lab4 Marks, Mini Project Marks.

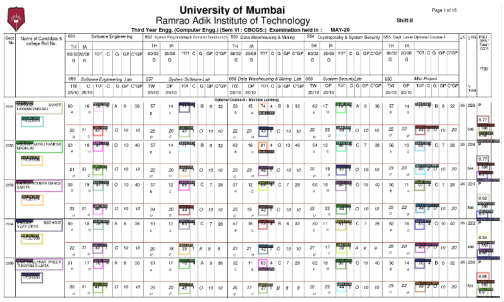

Figure 5. Detected Fields

Figure 5 shows the fields detected by the model for one single page of Exam Result Gazette PDF. Along with the detected fields and their respective labels the confidence percentage is also mentioned beside the boxes. While testing we have kept a threshold of 80% that means only the fields with confidence percentage greater than 80% are finally given as output.

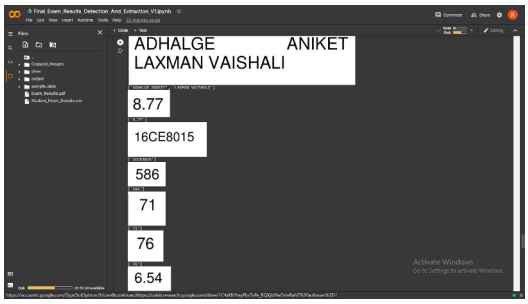

Figure 6. OCR Output

The detected fields from Exam Result Gazette Document are cropped with OpenCV library and are then passed to PyTesseract Optical Character Recognition System to extract the data as shown in Figure 6.

In this paper we have worked with Exam Result Gazette PDF consisting of 15 pages having total 71 student records. After giving the whole PDF to our system, after all processing, a CSV file is generated, which could then be downloaded and accessed locally as shown in Figure 7.

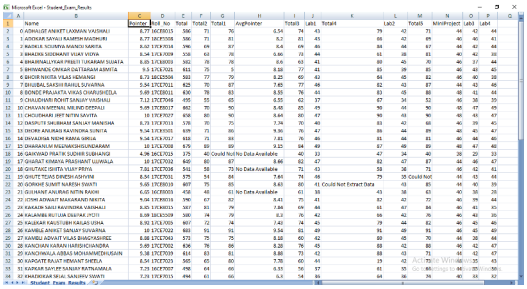

Figure 7. Output CSV File

The total fields that the model had to detect and recognize from the test set were around 1605 fields, out of which 1584 fields were detected and recognized by the model correctly. The metric that we used for evaluating our whole system consisting of detection and recognition part is Accuracy. Precision and Recall metrics have not been used because the system gives the complete extracted data as the output but if some of the digits in a detected field are wrong it will result in overall wrong prediction of the field. So this renders Precision and Recall as unsuitable evaluation metrics. The formula used to calculate overall accuracy after detection and recognition is given in Equation 1.

$$Accuracy = \frac{Total\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

The system has 98.69% accuracy. Name, Pointer, Roll Number, Lab1 Marks, Lab2 Marks, Lab4 Marks fields have an accuracy of 100%.

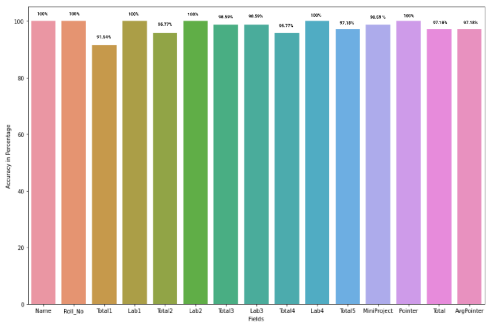

Figure 8. Accuracy Of Fields

Figure 8 shows the Accuracy of model across all the fields detected and recognized from the Exam Results Gazette PDF.

5 Conclusion And Future Work

Documents have valuable information locked in them, which could be utilized for fast and efficient searching and for business processing to get insights out of them, but one caveat is that it is largely a manual effort. A system that

can extract all this data automatically can significantly improve the efficiency of many corporate workflows.

Here, we have studied recent research papers and carried out a comprehensive literature survey based on the topic of structured data extraction from documents. We have trained and hyper-tuned a model for automatically detecting the values of required fields from the documents. We worked with the Exam Result Gazette Document. We got an overall accuracy of 98.69% on unseen Exam Results Gazette document. In future work, this system can be expanded to detect and recognize fields across various business domains.

References

[1] A. J. Sellen, R. H. Harper, "The Myth of the Paperless Office", MIT Press. (2003)

[2] B. Klein, S. Agne, A. Dengel, “Results of a Study on Invoice Reading Systems in Germany”, Springer. 451–462 (2004)

[3] K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask R-CNN", CVPR. (2018)

[4] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, "Aggregated Residual Transformations for Deep Neural Networks", CVPR. (2017)

[5] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature Pyramid Networks for Object Detection", CVPR. (2017)

[6] "Common Objects In Context", https://cocodataset.org, Accessed: 2020-12-10

[7] B. Majumder, N. Potti, S. Tata, J. Wendt, Q. Zhao, M. Najork, "Representation learning for information extraction from form-like documents", ACL. 6495–6504 (2020)

[8] T. Denk, C. Reisswig, "Bertgrid: Contextualized embedding for 2d document representation and understanding", NeurIPS. (2019)

[9] V. Sunder, A. Srinivasan, L. Vig, G. Shroff, R. Rahul, "One-shot information extraction from document images using neuro-deductive program synthesis", Neural-Symbolic Learning and Reasoning at IJCAI. (2019)

[10] B. Davis, B. Morse, S. Cohen, B. Price, C. Tensmeyer, "Deep visual template-free form parsing", ICDAR. (2019)

[11] S. Paliwal, V. D, R. Rahul, M. Sharma, L. Vig, "Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images", ICDAR. (2019)

[12] X. Holt, A. Chisholm, "Extracting structured data from invoices", ALTA. (2018)

[13] X. Zhao, E. Niu, Z. Wu, X. Wang, "Cutie: Learning to understand documents with convolutional universal text information extractor", CVPR. (2019)

[14] A. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, J. Faddoul, "Chargrid: Towards understanding 2d documents", EMNLP. (2018).

[15] V. D, R. Rahul, G. Sehgal, Swati, A. Chowdhury, M. Sharma, L. Vig, G. Shroff, A. Srinivasan, "Deep reader: Information extraction from document images via relation extraction and natural language." Asian Conference on Computer Vision, Springer. (2018)

[16] R. Palm, O. Winther, F. Laws, "Cloudscan-a configuration-free invoice analysis system using recurrent neural networks", ICDAR. **Vol. 1.** IEEE. (2017)